



한글 자모 구조를 활용한 3채널 LLM 워터마킹

Three-Channel LLM Watermarking via Hangul Jamo Decomposition

TEAM

강희지, 강재현, 박서현, 여운봉 | 연세대학교 컴퓨터과학과

컴퓨터종합설계 2025

CAPSTONE DESIGN PROJECT

1 연구 배경 및 문제점

기존 BPE 토크나이저의 한계

- 형태소 훠손: 한글의 의미적 최소 단위가 파괴됨
- 용량 부족: 단일 토큰당 워터마크 삽입 공간 협소
- 재토큰화 오류: 탐지 과정에서 정보 유실 발생

Conventional (BPE)

Low Capacity
인공 ##지능

Proposed (Jamo)

3x Capacity
ㄴ ㅡ ㅇ

제안 솔루션 (Solution)

3채널 자모 분리: 초/중/종성을 독립적인 정보 채널로 활용하여 삽입 용량 증대

Generation-time 개입: Logits 단계에서 확률을 편향시켜 품질 저하 없이 실시간 삽입

2 핵심 기술: 3채널 자모 분해

워터마크 삽입 과정 예시

Payload (메시지):

"ABC" → 67 68 69
→ 0100 0001 0100 0010...

TARGET: 01

이번 단계에 선택할 음소: 초성

후보 1: "인공지능"

Prob: 0.28 → 0.65

ㄴ

Index: 1

9 mod 4 = 01 (Match!)

후보 2: "학습"

Prob: 0.22

ㅅ

Index: 7

7 mod 4 = 11 ×

3 시스템 파이프라인

01
10

Payload Input

메시지를 이진 비트스트림으로 변환

ABC → 01000001...

██

Logits Prediction

LLM의 다음 토큰 확률 분포 예측

Top-K Candidates

☲

Jamo Decomposition

후보 토큰의 마지막 음절을 자모 단위로 분해

[초, 중, 종] × 3 Channels

⚡

Bit Matching & Biasing

Target 비트와 일치하는 토큰의 확률을 증폭

Logits[match] += 6

🕒

Watermarked Text

워터마크가 삽입된 텍스트 생성 완료

Output Generation

4 실험 결과

INPUT PAYLOAD

"ABC" → 01000001...

GENERATED TEXT (KOGPT2)

"이를 위해 우리는 인공지능에 주목하고, 인간 중심의 로봇·자동차... (중략)"

* 실제 생성 결과 반영

DETECTION RESULT

Extracted Bits: 01000001...

Recovered Msg: "ABC"

Bit Accuracy

99.8%

Capacity

3 bits/token

5 결론 및 기대효과

한국어 최적화

자모 구조를 보존하는 최초의 3채널 워터마킹

고용량 데이터

기준 대비 획기적인 정보 삽입량 확보

오픈소스 확장성

다양한 LLM 및 토크나이저에 즉시 적용 가능

본 기술은 AI 생성 콘텐츠의 저작권을 보호하고 불법 유포를 추적하는 기술로 활용될 수 있을 것입니다.



GitHub

프로젝트 데모 및 보고서 전문 확인